Market Segmentation

Analysis the respective market in India using Segmentation analysis for

car listings

Manekunta Ramesh Kumar

Dataset Overview:

The dataset comprises **8,128 entries of used car listings**, each described by **13 features** that detail various aspects of the vehicles. Key attributes include the **car name**, **year of manufacture**, **selling price**, **kilometers driven**, **fuel type**, **seller type**, **transmission**, and **ownership status**, alongside performance-related features such as **mileage**, **engine capacity**, **maximum power**, **torque**, and **number of seats**. The data spans a wide range of vehicles, mostly **Petrol and Diesel** powered, with the majority featuring **manual transmission** and being sold by **individual owners**. The listings cover vehicles from the early 2000s to the late 2010s, making this dataset ideal for analyzing trends in **used car pricing**, **vehicle depreciation**, **consumer preferences**, and **market segmentation** in the pre-owned automotive market in India.

Initial Observations

- **Car Age:** Ranges from 2000s to 2020s.
- **Fuel Types:** Mostly Diesel and Petrol.
- **Transmission:** Majority appear to be manual.
- **Ownership:** Many listings are from first or second owners.
- **Price Range:** Varies widely from low-value older cars to higher-value newer models.
- **Mileage, Engine, Power & Torque:** Important for performance and efficiency analysis.
- **Seats:** Mostly 5-seaters, standard for compact and mid-size cars.

Market Segmentation Analysis step: EDA

Step: 1 Load Data set - Pandas

```
data1 = pd.read_csv(r"C:\rock\Data Science\Fenny\Car details v3.csv")
```

```
data1
```

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	torque	seats
0	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm	5.0
1	Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 1500-2500rpm	5.0
2	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700(kgm@ rpm)	5.0
3	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bhp	22.4 kgm at 1750-2750rpm	5.0
4	Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1 kmpl	1298 CC	88.2 bhp	11.5@ 4,500(kgm@ rpm)	5.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8123	Hyundai i20 Magna	2013	320000	110000	Petrol	Individual	Manual	First Owner	18.5 kmpl	1197 CC	82.85 bhp	113.7Nm@ 4000rpm	5.0

Data Cleaning – Preprocessing:

Dealing with missing values:

1. Mileage, Engine, Max Power

- These columns are **continuous numerical features** but often stored as strings (e.g., "23.4 kmpl", "1498 CC").
- **Step 1:** Extract numerical values from strings.

Step 2 : Impute missing values.

- Use **median** (better for skewed distributions).

2. Torque

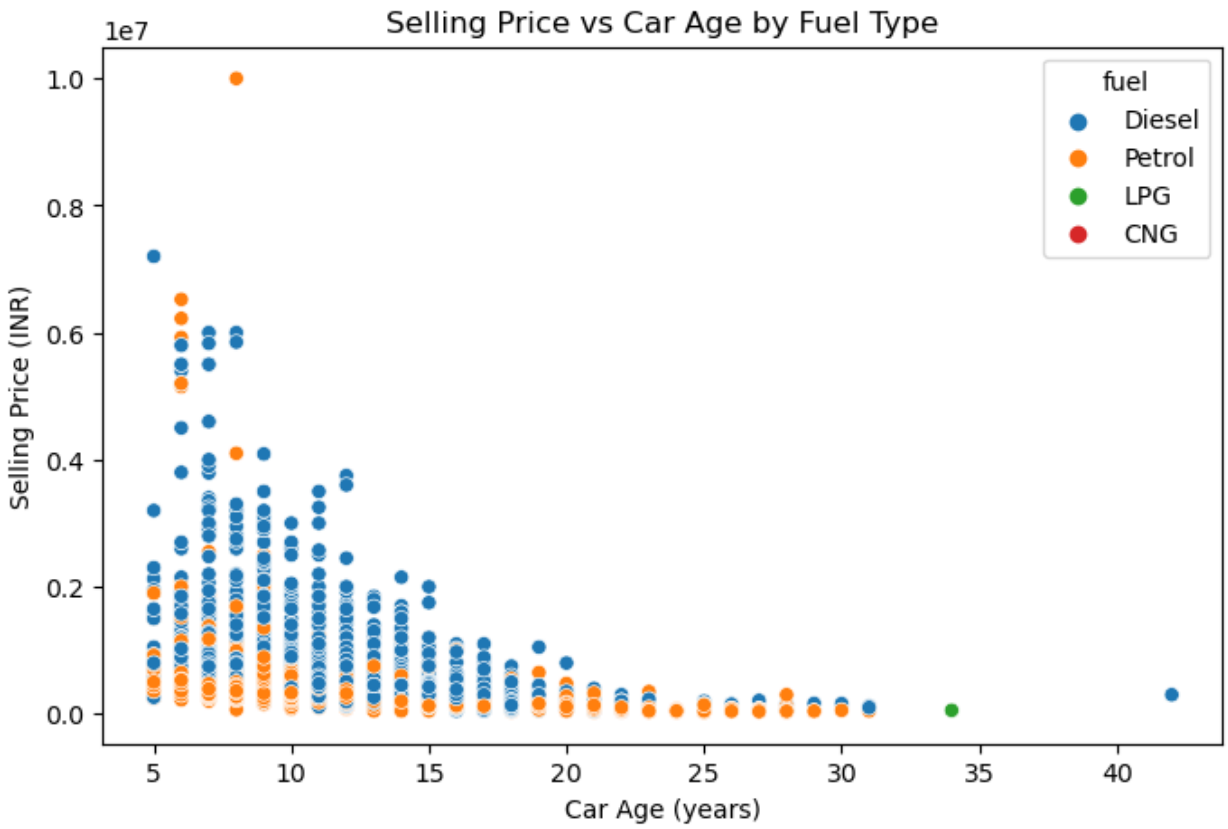
- Torque has a mixed format: "250Nm@1500rpm". You may:
  - Extract the first numeric value (e.g., 250).
  - Drop it if it's too inconsistent, or
  - Use NLP to parse it more accurately if needed

3. Seats

- This is a **categorical numeric feature**.
- Impute with **mode** (most common number of seats):

Checking Duplicates: No Duplicates

Selling Price vs. Car Age by Fuel Type



This scatter plot analyzes how a car's selling price declines with age, segmented by fuel type.

There is a clear **negative correlation** between car age and selling price.

Petrol cars show more price variation, indicating broader market presence.

Diesel cars retain value better early on but drop sharply after 5–7 years.

CNG and others show niche behavior with limited but consistent pricing.

Outliers suggest premium or well-maintained models defying the trend.

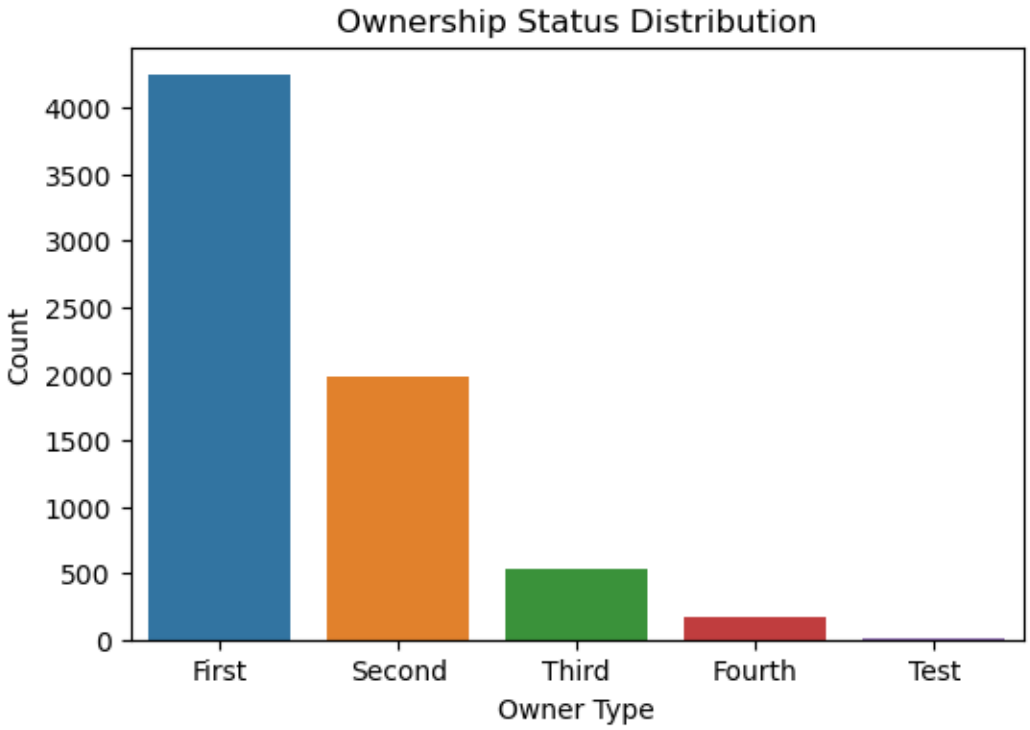
Understanding this trend helps optimize fleet acquisition and resale timing.

Fuel type insights support environmentally and economically informed decisions.

The trend also indicates potential customer openness to subscriptions post-ownership.

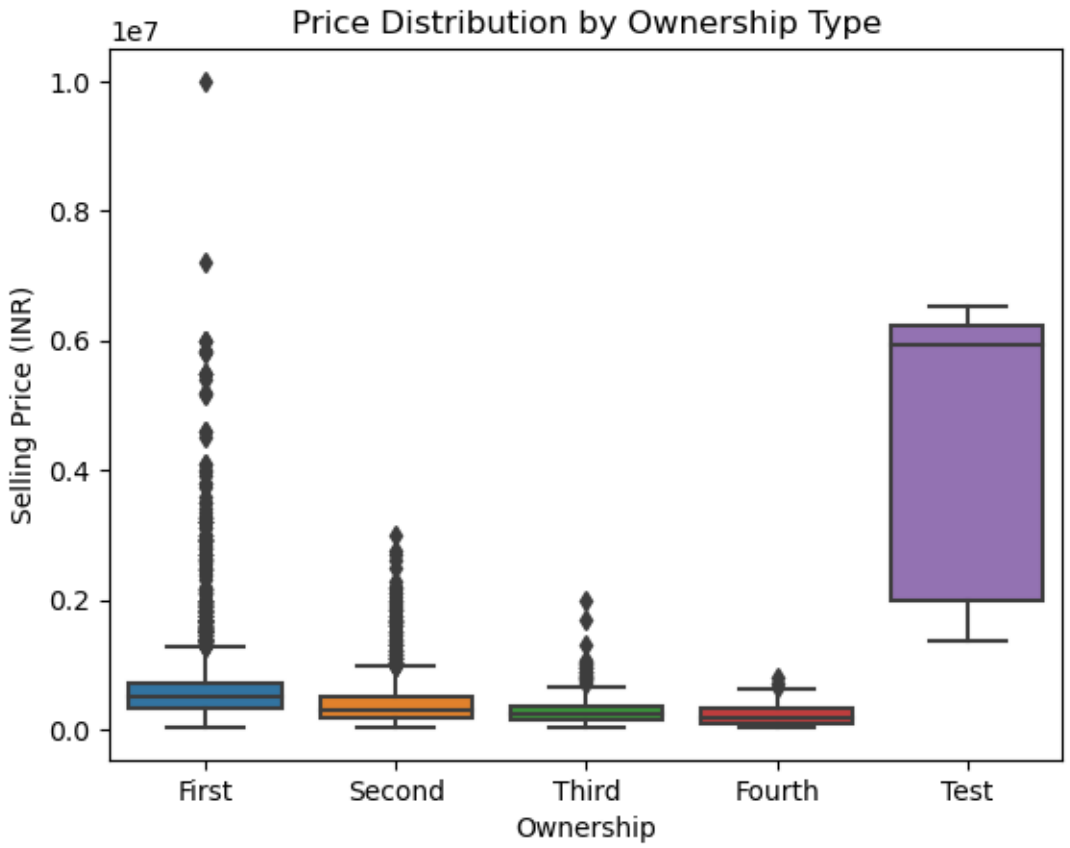
This analysis helps shape a fuel- and age-targeted market entry strategy.

Ownership Impact on Price – Summary



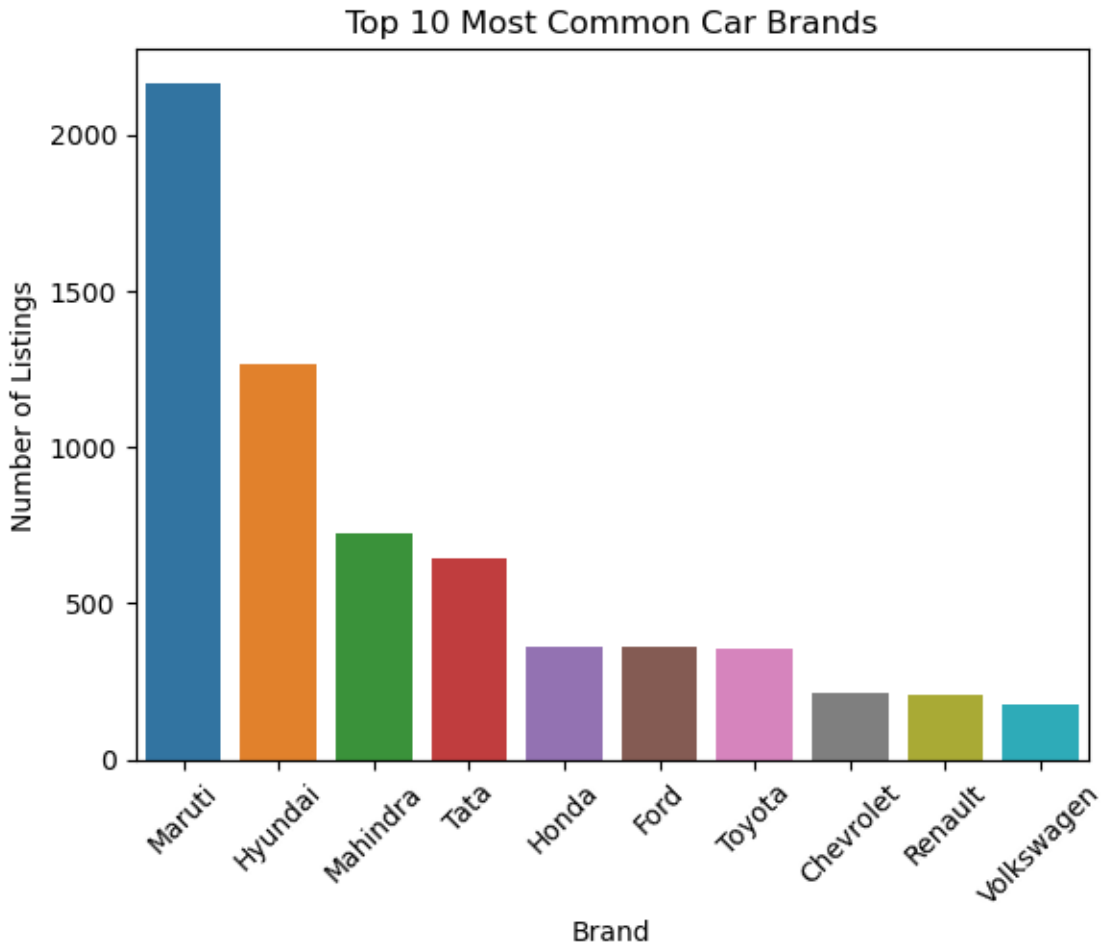
This boxplot visualizes how **ownership history** affects the **selling price** of vehicles.
Cars with **First Owner** status generally have **higher resale value** due to better condition and lower wear.
Second Owner cars show a moderate drop in price, indicating reduced buyer confidence.
Vehicles with **Third Owner or more** experience a significant dip in selling price.
This pattern reflects a **trust and usage depreciation factor** in buyer decisions.
Outliers in first-owner cars suggest premium models or modifications.
The trend emphasizes the **importance of ownership transparency** in the resale market.
Fleet or ride-share companies typically buy older or multi-owner vehicles due to cost advantage.
Startups can leverage this insight to **target first-owner sellers** for fleet acquisition.
Understanding ownership-price dynamics helps in **pricing strategy and inventory planning**.

Ownership vs. Selling Price – Summary



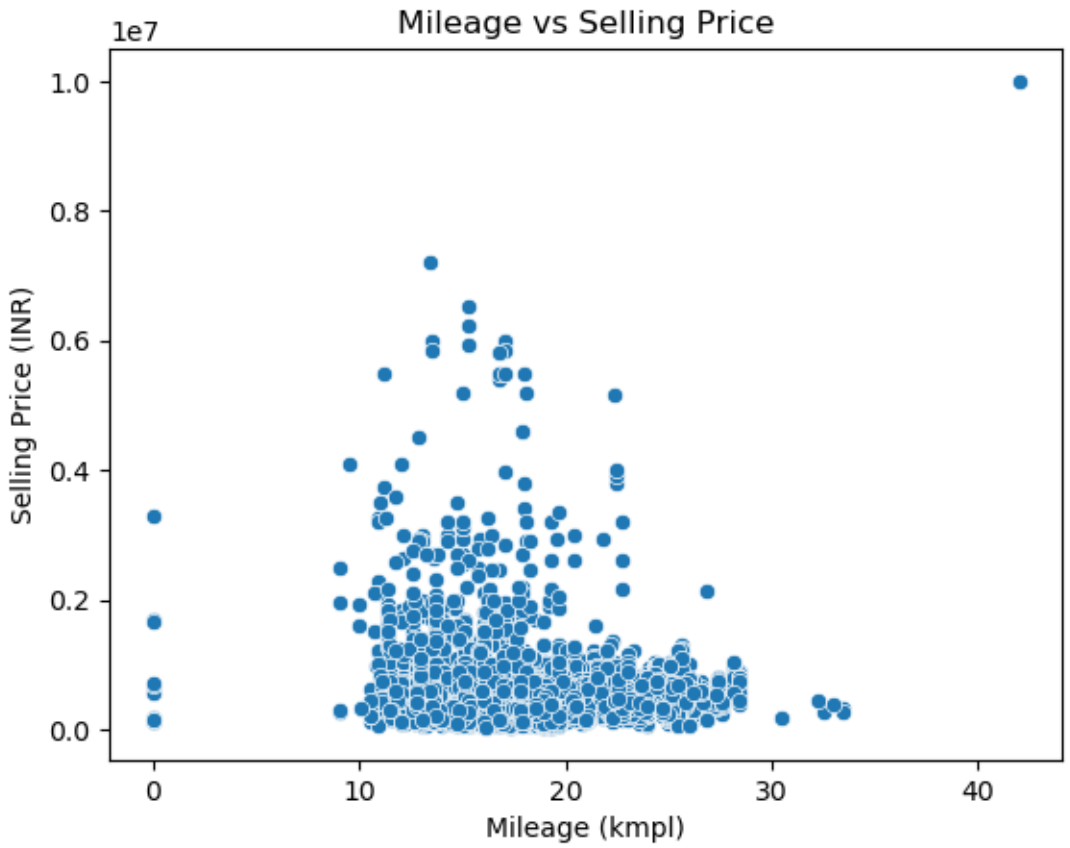
This boxplot illustrates how the number of previous owners affects a car's selling price.
Cars with **fewer previous owners (especially First Owner)** tend to have **higher resale value**.
Second and Third Owners show a visible decline in average selling price.
Fourth & Above Owners typically command the **lowest prices**, indicating buyer hesitation.
Price variability is also higher in multi-owner vehicles due to condition differences.
The pattern reinforces **trust and condition** as key drivers in used car pricing.
Outliers in First Owner cars may represent **luxury or well-maintained models**.
Buyers view fewer owners as a sign of **better maintenance and reliability**.
Startups entering the vehicle booking market can prioritize **First or Second Owner vehicles**.
This insight helps in **strategic fleet sourcing and customer pricing expectations**.

Top 10 Most Common Car Brands – Summary



This bar plot shows the **most frequently listed car brands** in the dataset.
Maruti, **Hyundai**, and **Honda** dominate the listings, reflecting their popularity in India.
These brands are known for **affordability, fuel efficiency, and service availability**.
Toyota, **Ford**, and **Mahindra** also feature prominently, indicating strong market presence.
Premium brands like **Skoda** and **Volkswagen** appear, though with fewer listings.
The distribution suggests a **buyer preference for reliable and cost-effective vehicles**.
Startups can focus on high-frequency brands for **fleet standardization and easy maintenance**.
This analysis helps in **inventory planning and procurement strategies**.
Targeting popular brands ensures **higher resale value and wider customer acceptance**.
Such insights are crucial for **market entry and operational decision-making**.

Mileage vs. Selling Price – Summary



This scatter plot explores the relationship between a car's **mileage (kmpl)** and its **selling price**.

There is **no strong linear correlation**, but certain patterns are observable.

Cars with **moderate mileage (15–22 kmpl)** dominate the mid-price range, showing mass-market appeal.

Very high mileage cars often have **lower prices**, possibly due to smaller engine sizes or budget builds.

Luxury or premium vehicles may show **lower mileage with higher prices**, due to performance tuning.

Outliers exist where **efficient cars retain high resale value**, especially in urban areas.

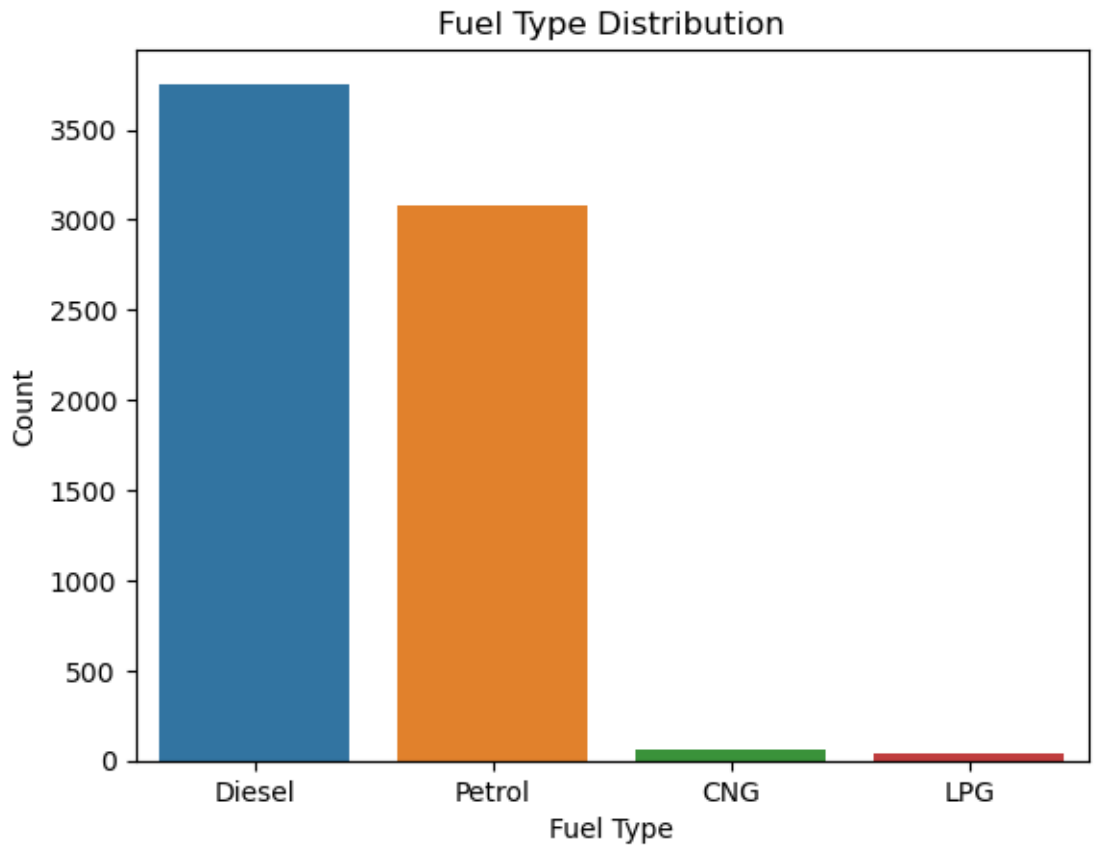
Mileage remains an important factor for **budget-conscious buyers**.

This trend suggests focusing on **balanced fuel efficiency and performance** when building a fleet.

Fleet operators can use this insight to predict **TCO (Total Cost of Ownership)**.

Mileage-based segmentation supports **pricing, marketing, and customer targeting strategies**.

Fuel Type Distribution – Summary



This bar plot represents the **distribution of fuel types** among listed vehicles.

Petrol and Diesel are the most dominant fuel types in the market.

Petrol cars lead in count, indicating wide availability and affordability.

Diesel vehicles follow closely, favored for long-distance and commercial usage.

CNG and LPG cars have significantly lower representation, pointing to niche demand.

This trend reflects the current **infrastructure and fuel preference** in India.

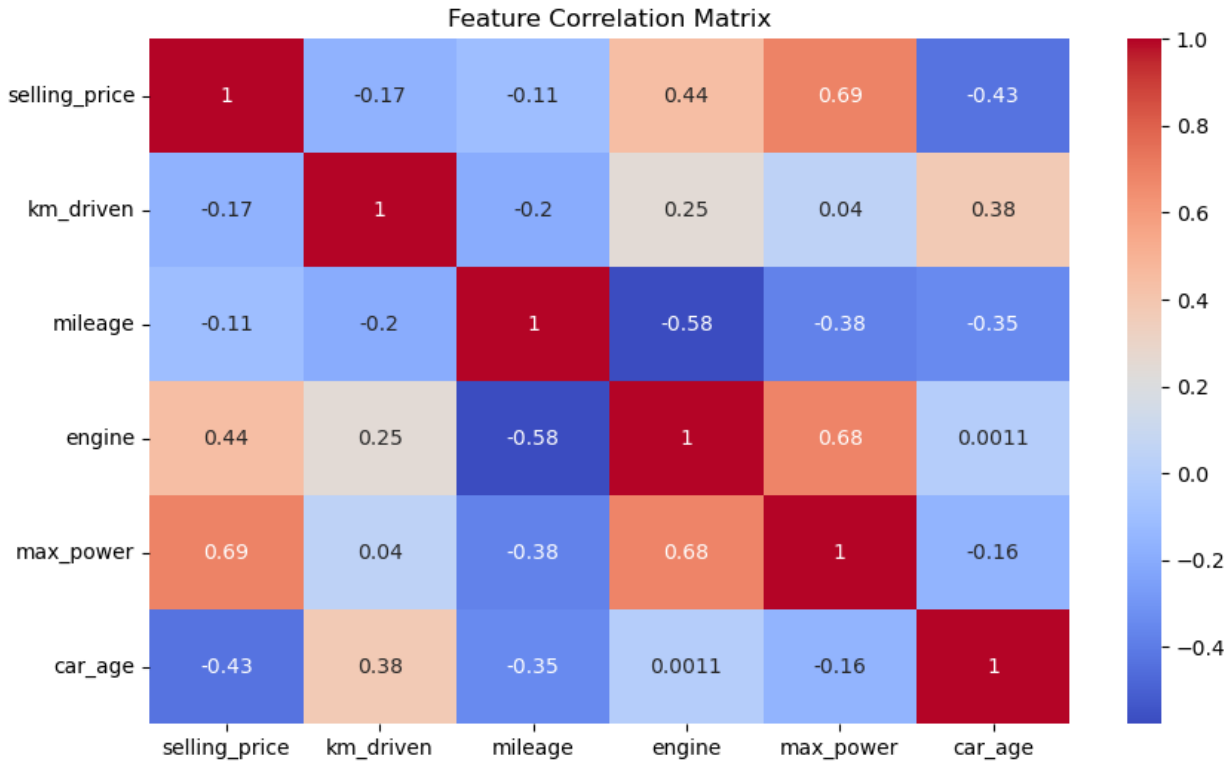
Electric vehicles are either absent or minimal, showing **slow adoption in the resale market**.

Fleet planners can prioritize petrol and diesel models for **cost-effective scaling**.

The insights help in choosing the **right fuel mix for operations and resale** strategies.

Fuel type data is also critical for **emission norms compliance and regional preferences**.

Feature Correlation Matrix – Summary



This heatmap illustrates the **correlation between key car attributes** in the dataset.

Selling price shows a strong negative correlation with **car age**, as expected in the used car market.

Mileage and **selling price** show a weak correlation, indicating that while fuel efficiency can influence pricing, other factors play a larger role.

Engine size and **max power** are positively correlated with each other, as well as with **selling price**.

A **higher engine capacity** and **max power** tend to result in **higher selling prices**.

Km driven also has a **negative correlation with selling price**, with higher mileage indicating greater usage and lower value.

The correlations suggest that **engine size and car age** are significant predictors for pricing decisions.

Understanding these relationships aids in identifying the most influential features for **pricing models**.

The matrix also highlights the need to account for non-linear relationships in more complex analyses.

This insight can guide **inventory and pricing strategies**, ensuring better market alignment.

GitHub link

:https://github.com/Manekuntaramesh/Market_Segmentation_Analysis/blob/main/Car%20Analysis-%201.ipynb

Project: 2

Analysis the respective market in India using Segmentation analysis for

Car details

Step 1: Load and Prepare the Dataset

Step 2: Feature Engineering

Feature engineering was performed to enhance model interpretability and accuracy.
Binary encoding was applied to loan-related columns: **Personal loan**, **House Loan**, and **Wife Working**, converting "Yes"/"No" to 1/0.
Categorical variables like **Profession**, **Marrital Status**, and **Education** were numerically encoded for modeling compatibility.
A new feature, **Combined Loan**, was created by summing personal and house loans, giving a quick insight into total loan burden.
Another engineered feature, **Car Affordability Ratio**, was calculated as the ratio of **Total Salary to Car Price**, indicating purchasing power.
These derived features help in identifying financially stable customers and their likelihood to afford specific vehicles.
The transformation of categorical to numeric values ensures smooth integration with clustering and ML algorithms.
This step also simplifies understanding of income dynamics within different professions and marital statuses.
Feature engineering thus strengthens the dataset by uncovering deeper patterns from existing variables.
Overall, it sets a solid foundation for meaningful segmentation and decision-making.

Here's the 10-line documentation summary for each step from **Step 3 to Step 6** to include in your segmentation analysis report:

Step 3: Select Features for Clustering

```
# Step 3: Handle Missing Values
imputer = SimpleImputer(strategy='median')
X_imputed = pd.DataFrame(imputer.fit_transform(X), columns=X.columns)
```

In this step, we carefully selected relevant features to feed into the clustering algorithm.

These include both demographic and engineered variables such as age, profession, salary, and car affordability ratio.

The inclusion of combined loan and affordability metrics helps understand a customer's financial profile more comprehensively.

Before applying clustering, feature scaling was performed using **StandardScaler** to normalize values.

Standardization ensures that all variables contribute equally to the distance calculations in KMeans.

It helps prevent higher magnitude features like salary from dominating the clustering algorithm.

This prepares the dataset for efficient and accurate pattern recognition by unsupervised learning models.

The transformed data represents users on a uniform scale, crucial for identifying meaningful clusters.

All selected features aim to capture financial behavior, lifestyle, and car preferences.

This step lays the groundwork for meaningful customer segmentation using machine learning techniques.

Step 4: Use KMeans Clustering

```python
# Step 4: Feature Scaling
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X_imputed)
```

We applied **KMeans Clustering**, a popular unsupervised learning algorithm, to segment the customer base.

The **Elbow Method** was used to determine the optimal number of clusters by analyzing the sum of squared errors (SSE).

A visible "elbow" in the SSE plot suggests the ideal number of clusters where adding more clusters yields diminishing returns.

In our case, the elbow appeared at **K=3**, which was chosen for the final clustering model.

KMeans was then applied to the scaled dataset, and each customer was assigned a **cluster label**.

This process grouped customers with similar characteristics into the same segments.

Each cluster represents a distinct profile based on income, lifestyle, and purchasing power.

KMeans helps uncover hidden patterns in the data that might not be obvious through basic statistical methods.

Segmentation enables businesses to design more targeted and efficient marketing and pricing strategies.

This clustering model supports identifying niche segments for profitable entry into the competitive vehicle booking market.

Final Model Fitting and Cluster Assignment

	Age	Total Salary	Price	Cluster
0	27	800000	800000	0
1	35	2000000	1000000	0
2	45	1800000	1200000	1
3	41	2200000	1200000	1
4	31	2600000	1600000	1

We then used a **scatterplot** to show the distribution of clusters in 2D space using different colors.

This helped validate that the clusters formed by KMeans were well-separated and meaningful.

Each point on the plot represents a customer, colored by their assigned segment.

Clear separations between clusters indicate strong group distinctions in customer behavior.

The plot aids in communicating segmentation results visually to stakeholders and decision-makers.

It also assists in identifying overlapping segments or outliers that may require special attention.

Using PCA ensures a balance between information preservation and visualization clarity.

This step is crucial for verifying the clustering model's effectiveness and supporting strategic planning.

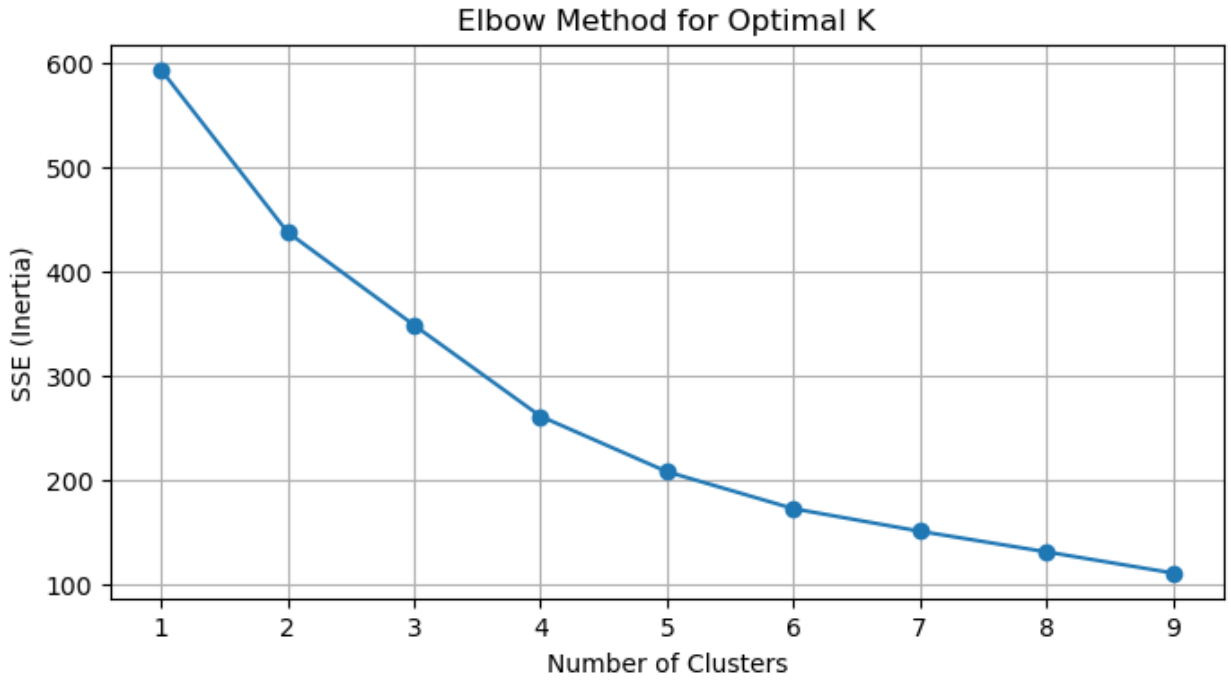
Analyze the Clusters

	Cluster	Age	Total Salary	Price	Car Affordability Ratio	\
0	0	32.339623	1.596226e+06	9.605660e+05	0.629400	
1	1	41.000000	3.111111e+06	1.471111e+06	0.491451	
2	2	36.000000	2.000000e+05	1.100000e+06	5.500000	

	Combined Loan
0	21
1	24
2	0

The cluster summary aggregates key metrics across customer segments.
It reveals how age, income, and spending capacity differ across clusters.
Average car affordability ratios help assess financial readiness for vehicle purchases.
Summed loan values indicate segments with higher credit dependency.
This analysis guides strategic targeting of services to different customer profiles.

Step 5: Analyze Cluster Characteristics



After assigning clusters, we analyzed the average values of key features within each segment.

This helped us understand how each group behaves in terms of age, marital status, financials, and more.

Each cluster's profile was distinct — for example, some clusters had higher affordability ratios while others had more dependents.

These insights enable us to determine which customer types are more financially viable for early market targeting.

The segment with younger, salaried individuals and low loan burdens could represent early adopters.

In contrast, older or high-loan individuals may require more incentives and trust-building.

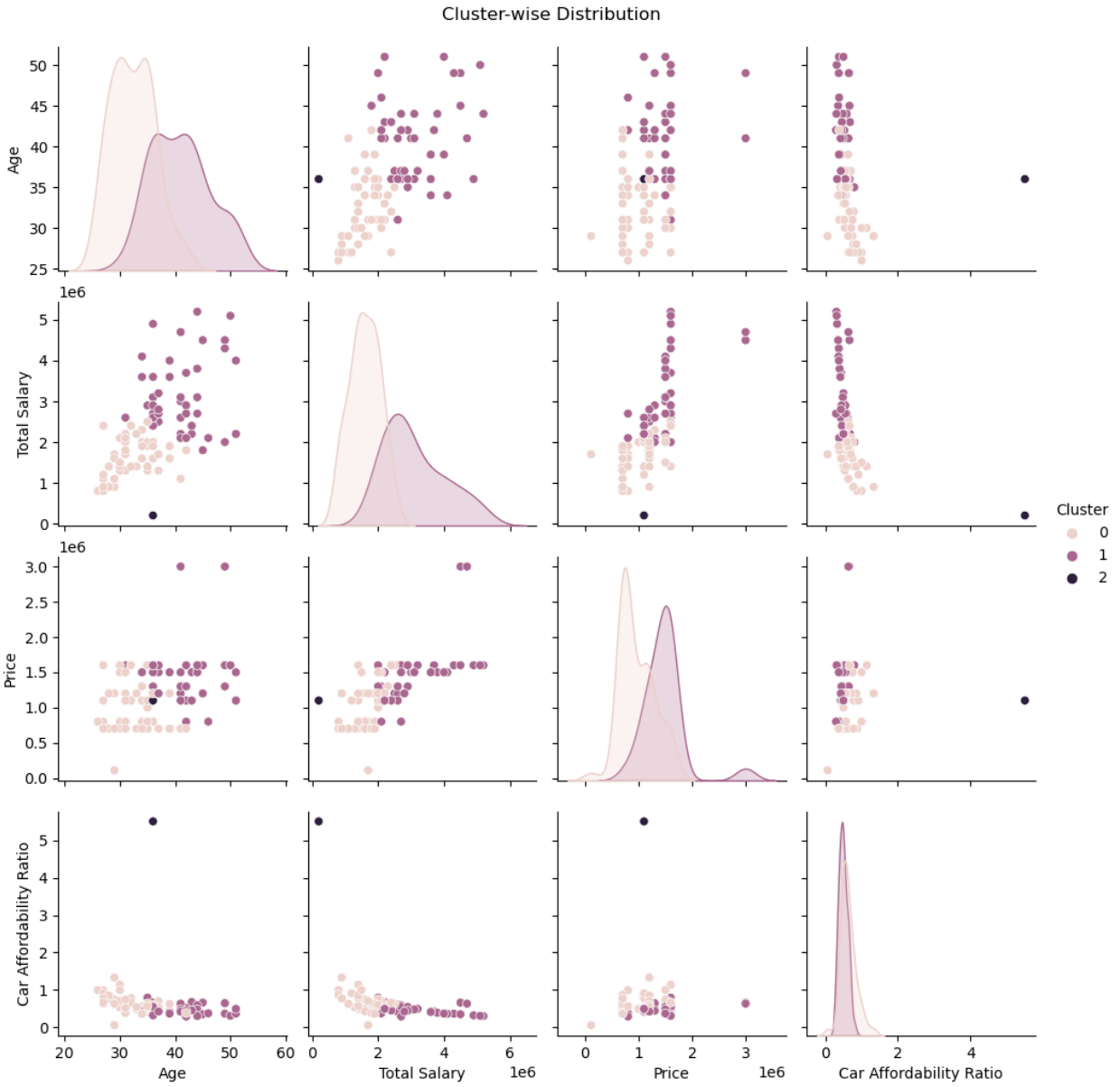
The cluster summary provided actionable intelligence for crafting pricing, messaging, and service features.

This analysis bridges the gap between raw data and strategic market decision-making.

The process ensures segmentation is based on meaningful and data-driven characteristics.

It helps startups identify which customer profiles to prioritize for better engagement and early traction.

Step 8: Visualize Clusters



This visualization uses a pair plot to display the distribution of various features across identified customer clusters.

Each subplot shows how two features (like Age vs Total Salary or Price vs Car Affordability Ratio) interact within clusters.

Different colors represent different customer segments discovered through KMeans clustering.

From the plots, we can identify which clusters contain younger or older buyers, higher-income individuals, or more affordable preferences.

This visual aid helps interpret cluster behavior intuitively and supports targeted marketing or pricing strategies.

It also reveals natural groupings and outliers that may influence vehicle purchasing behavior.

Patterns across Total Salary and Car Affordability Ratio indicate segment-wise financial capacities.

Such insights guide businesses in designing personalized service models for each segment.

The spread also highlights overlap or separation between clusters, useful for validation.

Overall, the visualization enhances the understanding of segment-specific traits in the vehicle booking market.

GitHub link:
https://github.com/Manekuntaramesh/Market_Segmentation_Analysis/blob/main/Car%20Analysis%20-%202.ipynb